

Joint Analysis of Longitudinal and Recurrent Event Outcomes

Elizabeth H. Slate

Dept. of Biostatistics, Bioinformatics, and Epidemiology
Medical University of South Carolina
Charleston, SC 29425 USA
slateeh@musc.edu

Edsel A. Peña

Department of Statistics
University of South Carolina
Columbia, SC 29208 USA
pena@stat.sc.edu

Abstract

We consider modeling the joint outcomes of a recurrent event process and an associated longitudinal biomarker. Example studies where such data may arise include those monitoring repeated heart attacks and related markers such as cholesterol levels or inflammation measures, or studies assessing association between markers for diabetes control and recurrent oral complications. In this talk, we describe a latent class model in which the event process and longitudinal outcome are conditionally independent given the latent class. The model for the recurrent event accommodates the effect of accumulating event occurrences on the subject, as well as the effects of interventions performed after each event occurrence. We discuss inference for this joint model and illustrate behavior under varying simulation scenarios and by application to biomedical data.

1 Introduction

In a wide variety of settings, the event of major interest is recurrent. Common examples of such events in the biomedical setting are epileptic seizures, occurrence of tumors, and hospitalization due to a chronic disease. In engineering and reliability settings, recurrent events are the failure of a computer software, a nuclear power plant meltdown, or the failure of a major subsystem of a spacecraft. In actuarial situations, automobile and non-life insurance claims are recurrent events, while personnel absenteeism and criminal offenses are recurring events with a sociological flavor. A drop of 15% of the Dow Jones industrial average during a trading day is a recurrent event with economic connotations, and international conflicts or formal accords could be recurrent events in the political arena.

In many of these situations, there is a longitudinal variable associated with the event occurrence rate that can be evaluated repeatedly for units under study. Examples in the biomedical setting include prostate-specific antigen (PSA) as a longitudinal biomarker associated with recurrence of prostate cancer, cholesterol level, blood pressure, or ventricular hypertrophy as markers related to the occurrence of heart attacks, and liver enzyme measurements as markers for repeated hospitalization due to hepatitis. Including these longitudinal markers in the modeling may improve the predictive ability of the model concerning event occurrences. Alternatively, these markers may themselves be a focus of interest, and the data on the event occurrences may enhance the modelling and understanding of the evolution of the longitudinal marker. Because the marker is itself a genuine outcome variable, there is potential benefit in terms of predictive ability and ease of interpretation to model the marker and recurrent event as jointly varying response variables. Such a model was considered by Lin, Turnbull, McCulloch, and Slate (2002) in the single event setting.

Lin et al. (2002) postulated a latent class model wherein, conditionally on the latent class to which a unit belongs, the longitudinal marker trajectories and the occurrence of the event are stochastically independent. In effect, the association between the longitudinal marker process and the event process is entirely encoded in the latent class to which the unit belongs. However, because the latent class membership is unobserved, marginalization with respect to the possible latent classes yields the association between the marker and the event occurrence. They demonstrated the viability of this latent class model in the single-event case by applying it to a real data set with the longitudinal marker being the PSA level, and ascertained some properties of the inferential aspects for the model through computer simulations.

We propose extending the work of Lin et al. (2002) to accommodate joint modelling of recurrent event and longitudinal marker outcomes. In the analysis of inter-event times of recurrent events, it is important to account for an informative stopping time and informative censoring that arise because of a sum-quota accrual scheme: because the unit is typically observed over a random period, the frequency of event occurrence is determined by the successive inter-event times, and the censoring time following the last observed event depends upon the preceding inter-event times. Peña, Strawderman, and Hollander (2001) discussed the impact of these issues for nonparametric estimation of the inter-event time distribution in a renewal model.

Peña and Hollander (2004) developed a very general model for recurrent events, which, in addition to covariate effects, incorporated the effects of the past number of event occurrences, the impact of interventions performed following event occurrences, and frailties. Inferential procedures were developed for this model by Peña, Slate, and Gonzalez (2003).

In this paper, we build upon the appealing properties of the latent class model in Lin et al. (2002) by incorporating the general recurrent event model of Peña and Hollander (2004) to arrive at a latent class model for joint recurrent event and longitudinal marker outcomes. The model also takes into account other relevant covariates, it models the possible weakening (or strengthening) effect of accumulating event occurrences, it models the effect of interventions that are performed after each event occurrence through the effective time that governs a baseline hazard, and finally also incorporates the effects of unobserved frailty variables. This new class of models is described in Section 2. The remainder of the paper discusses some properties of the models, presents estimation methods based on the EM algorithm (Dempster, Laird, and Rubin 1977), evaluates these methods through simulation studies, and illustrates their application.

2 Proposed Class of Models

In this section, we introduce notation and describe the proposed class of models. We begin by considering unit (or subject) i among n units. This unit will be under observation over the period $[0, \tau_i]$, where τ_i is a, possibly random, termination time, which could be induced by an administrative constraint.

2.1 Latent Class Submodel

Following Lin, et al. (2002), it will be postulated that the subject belongs to one of K latent classes, where, at this stage, K will be assumed known. This latent class membership will be represented by a $K \times 1$ multinomial vector $\mathbf{C}_i = (C_{i1}, C_{i2}, \dots, C_{iK})^T$ with the components defined via

$$C_{ik} = \begin{cases} 1 & \text{if } i\text{th unit is in } k\text{th latent class} \\ 0 & \text{otherwise} \end{cases},$$

and the value of \mathbf{C}_i unobserved. The probability that the unit belongs to the k th latent class will depend on a subject-specific $m \times 1$ covariate vector $\mathbf{V}_i = (V_{i1}, V_{i2}, \dots, V_{im})^T$ and modelled according to a logit function. Thus, with $\pi_{ik} = \Pr\{C_{ik} = 1 | \mathbf{V}_i\}$, it is postulated that

$$\pi_{ik} = \frac{\exp\{\mathbf{V}_i^T \boldsymbol{\eta}_k\}}{\sum_{j=1}^K \exp\{\mathbf{V}_i^T \boldsymbol{\eta}_j\}}, \quad k = 1, 2, \dots, K, \quad (1)$$

with the identifiability restriction $\boldsymbol{\eta}_1 = \mathbf{0}$, and where $\boldsymbol{\eta}_k$ is an $m \times 1$ vector of parameters for the k th class.

2.2 Conditional Longitudinal Marker Submodel

To model the longitudinal marker component, for this subject, we denote by $\{Y_i(s), s \geq 0\}$ the longitudinal marker of interest, and this marker will be observed on the (calendar) times $s_{i1} < s_{i2} < \dots < s_{im_i}$. We denote by $\mathbf{s}_i = (s_{i1}, s_{i2}, \dots, s_{im_i})$. For this unit, we will therefore have the $m_i \times 1$ vector of values of the marker given by $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{im_i})^T$ where $Y_{ij} = Y(s_{ij})$. At each calendar time at which the marker is observed, the value of a time-dependent $p \times 1$ fixed covariate vector $\{\mathbf{X}_i(s), s \geq 0\}$ will also be observed, so for this unit there will be an $m_i \times p$ matrix given by $\mathbf{X}_i = [\mathbf{X}_{i1}, \mathbf{X}_{i2}, \dots, \mathbf{X}_{ip}]$, where $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, \dots, X_{ijm_i})^T$ is

the vector of observed values of the j th covariate at the m_i time points. Furthermore, covariates for random effects and for class-specific effects are denoted respectively by the $m_i \times q_1$ and $m_i \times q_2$ matrices given by

$$\mathbf{Z}_i = [\mathbf{Z}_{i1}, \mathbf{Z}_{i2}, \dots, \mathbf{Z}_{i1_1}] \quad \text{and} \quad \mathbf{W}_i = [\mathbf{W}_{i1}, \mathbf{W}_{i2}, \dots, \mathbf{W}_{i1_1}],$$

where $\mathbf{Z}_{ij} = (Z_{ij1}, Z_{ij2}, \dots, Z_{ijm_i})^T$ and $\mathbf{W} = (W_{ij1}, W_{ij2}, \dots, W_{ijm_i})^T$. Latent class-specific parameters will be in the $q_2 \times K$ matrix $\mathbf{M} = [\mu_1, \mu_2, \dots, \mu_K]$ with μ_k being a $q_2 \times 1$ vector containing parameters specific to the k th latent class.

The stochastic model for the longitudinal marker, *conditionally* on the latent class membership vector \mathbf{C}_i , is specified by a linear mixed model given by

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \mathbf{W}_i(\mathbf{M}\mathbf{C}_i) + \epsilon_i, \quad (2)$$

where β is a $p \times 1$ vector of regression parameters, \mathbf{b}_i is a $q_1 \times 1$ random vector which is multinormally distributed with mean vector $\mathbf{0}$ and covariance matrix \mathbf{D} , and the error component ϵ_i is an $m_i \times 1$ vector with mean $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}_{m_i}$, with \mathbf{I}_a being an identity matrix of dimension a . Furthermore, it is assumed that \mathbf{b}_i and ϵ_i are stochastically independent, and in addition, given the latent classes and the random effects, the times $\{s_{ij}, j = 1, 2, \dots, m_i\}$ at which the marker and the covariates are observed are noninformative.

2.3 Conditional Recurrent Event Submodel

Next, we describe the conditional model for the recurrent event process. For the i th unit, we denote by $S_{i1} < S_{i2} < \dots$ the successive calendar times of event occurrences, and with $\mathcal{Z}_+ = \{0, 1, 2, \dots\}$, define $K_i = \max\{j \in \mathcal{Z}_+ : S_{ij} \leq \tau_i\}$ to be the number of event occurrences over the period of observation. With $T_{ij} = S_{ij} - S_{ij-1}, j = 1, 2, \dots$ where $S_{i0} = 0$, the data available from monitoring the occurrence of the recurrent event are

$$(K_i, \tau_i, \{T_{ij}, j = 1, 2, \dots, K_i\}, \tau_i - S_{iK_i}),$$

where we note that $\tau_i - S_{iK_i}$ is the right-censored value associated with T_{iK_i+1} . Aside from the above data, we may also observe a possibly time-dependent $q_3 \times 1$ covariate vector process $\{\mathbf{x}_i(s), s \in [0, \tau_i]\}$, with some of these components of this vector process possibly coinciding with some of the covariates observed when monitoring the longitudinal marker or with covariates governing the class membership model (1).

Following Peña and Hollander (2004), stochastic process notation facilitates model and inferential development. Towards this end, define the (calendar-time) counting process $\{N_i^\dagger(s), s \geq 0\}$ and at-risk process $\{R_i^\dagger(s), s \geq 0\}$ according to

$$N_i^\dagger(s) = \sum_{j=1}^{\infty} I\{S_{ij} \leq s\} \quad \text{and} \quad R_i^\dagger(s) = I\{\tau_i \geq s\},$$

where $I\{A\}$ is the indicator function of an event A . With these processes, we may then represent the observable data according to

$$\{(N_i^\dagger(s), R_i^\dagger(s), \mathbf{x}_i(s)), s \geq 0\}. \quad (3)$$

Let $\{\mathcal{F}_{is}, s \geq 0\}$ be the natural filtration generated by the data in (3); then $R_i^\dagger(s), s \geq 0$ and $\mathbf{x}_i(s), s \geq 0$ are predictable processes with respect to this filtration. To complete the submodel for the recurrent event process, we introduce an *observable* predictable, nonnegative, piecewise differentiable, and piecewise nondecreasing process $\{\mathcal{E}_i(s), s \geq 0\}$, referred to as the *effective age* process, and denote by $\omega_i, i = 1, 2, \dots, n$, an unobservable frailty variable associated with the i th unit. The recurrent event model is then given by

$$\Pr\{dN_i^\dagger(s) = 1 | \mathcal{F}_{is-}, C_{ik} = 1, \omega_i\} = \omega_i R_i^\dagger(s) \lambda_{0k}[\mathcal{E}_i(s)] \rho[N_i^\dagger(s-); \alpha_k] \psi[\mathbf{x}_i^T(s) \gamma_k] ds. \quad (4)$$

In (4), $\lambda_{0k}(\cdot), k = 1, 2, \dots, K$, are unknown baseline hazard functions associated with the latent classes, $\rho(\cdot; \alpha)$ is a nonnegative function of known form with $\rho_k(0) = 1$, while $\psi(\cdot)$ is a nonnegative link function of

known form, usually taken to be $\psi(u) = \exp(u)$. The frailty variable ω_i will be assumed in this paper to be gamma distributed with the same shape and scale parameters to make the model identifiable, that is, $\omega_i \sim \mathcal{G}(\theta, \theta)$ where $\mathcal{G}(a, b)$ is a gamma distribution with mean a/b and variance a/b^2 .

As motivated in Peña and Hollander (2004), the effective age process $\mathcal{E}_i(\cdot)$ represents the effect of interventions (“repairs” in the engineering and reliability context) after each event occurrence. One example is *perfect repair*, $\mathcal{E}_i(s) = s - S_{iN_i^\dagger(s-)}$, for which, after an event occurrence, the effective age of the unit governing the baseline hazard returns to zero. Another example is *minimal repair*, $\mathcal{E}_i(s) = s$, where the unit restarts at an age equal to the age just before the failure. Peña and Hollander (2004) provide other forms of this effective age process and discuss how it accommodates the varied proposals in the literature.

The function $\rho(\cdot; \alpha)$ in (4) encodes the effect of accumulating event occurrences on the unit. If an increasing number of event occurrences leads to a weakening of the unit, such as an increasing number of non-fatal heart attacks, then this function will be non-decreasing; whereas if more event occurrences lead to improvements in the unit, such as the discovery of bugs in a computer software, then this will be a non-increasing function. A simple form for this function could be $\rho(j; \alpha) = \alpha^j, j \in \mathcal{Z}_+$. The $\psi(\cdot)$ function in (4) represents the link between the event occurrences and relevant covariates.

2.4 Conditional Independence

The longitudinal marker and recurrent event process are joined through the assumption of conditional independence given the latent class membership. Additionally, the n units under study are assumed to be independent.

Remainder of talk

By considering the latent class membership and frailty values as “missing data,” estimation methods follow from the EM algorithm of Dempster, Laird, and Rubin (1977). These methods are developed for parametric and nonparametric specification of the baseline hazard functions in (4), and the properties of the resulting estimators are evaluated through simulation studies. The model is then illustrated by application to real data.

Acknowledgements

E. Slate acknowledges NIH Grant CA077789, NIH COBRE Grant RR17696, and DAMD Grant 17-02-1-0138. E. Peña acknowledges NSF Grant DMS 0102870, NIH Grant GM056182, and NIH COBRE Grant RR17698. Both authors also thank the USC/MUSC Collaborative Research Program.

References

- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* 39, 1–38.
- Lin, H., B. W. Turnbull, C. E. McCulloch, and E. H. Slate (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association* 97(457), 53–65.
- Peña, E. and M. Hollander (2004). *Mathematical Reliability: An Expository Perspective*, Chapter 6. Models for Recurrent Events in Reliability and Survival Analysis, pp. 105–123. Kluwer Academic Publishers.
- Peña, E., E. Slate, and J. Gonzalez (2003). Semiparametric inference for a general class of models for recurrent events. Technical Report 214, Department of Statistics, University of South Carolina.
- Peña, E. A., R. L. Strawderman, and M. Hollander (2001). Nonparametric estimation with recurrent event data. *Journal of the American Statistical Association* 96(456), 1299–1315.